Domain Adaptive Knowledge Distillation for Driving Scene Semantic Segmentation

Divya Kothandaraman   Athira Nambiar   Anurag Mittal

ramandivya27@yahoo.in, {anambiar,amittal}@cse.iitm.ac.in

Indian Institute of Technology Madras, India

Abstract

Practical autonomous driving systems face two crucial challenges: memory constraints and domain gap issues. In this paper, we present a novel approach to learn domain adaptive knowledge in models with limited memory, thus bestowing the model with the ability to deal with these issues in a comprehensive manner. We term this as "Domain Adaptive Knowledge Distillation" and address the same in the context of unsupervised domain-adaptive semantic segmentation by proposing a multi-level distillation strategy to effectively distil knowledge at different levels. Further, we introduce a novel cross entropy loss that leverages pseudo labels from the teacher. These pseudo teacher labels play a multifaceted role towards: (i) knowledge distillation from the teacher network to the student network & (ii) serving as a proxy for the ground truth for target domain images, where the problem is completely unsupervised. We introduce four paradigms for distilling domain adaptive knowledge and carry out extensive experiments and ablation studies on real-to-real as well as synthetic-to-real scenarios. Our experiments demonstrate the profound success of our proposed method.

1. Introduction

Semantic segmentation is one of the fundamental tasks in perception for autonomous vehicles. The goal of semantic segmentation is to assign a label (car, tree, sky, etc.) to every pixel in the image. This information is quite relevant in the decision making of autonomous vehicles, as it provides information regarding shape/location of objects and the background. Recently, deep neural networks have imparted immense progress in this field. Some benchmark architectures include Fully Connected Networks [22], Dilated Residual Networks [43], and DeepLab [5], PSPNet [46], and Gated-scnn [34].

Despite these advancements, there are some practical challenges that cause a hindrance to even the best perform-

ing models. One such major drawback is the ***"domain-gap" issue***. It particularly exists in 'in-the-wild' scenarios, where neural networks trained on one domain won't generalize well on other domains, due to large domain gaps between the source (on which the neural network was trained) and the target (test) domains. This is mainly due to significant changes in the illumination, weather conditions and seasonal changes across domains. Additionally, collecting ground-truth (GT) annotations for different domains is a laborious and expensive task. This calls for an approach, that is capable of leveraging existing ground truth information in some domains to improve performance in other domains in an unsupervised manner. To this end, Unsupervised Domain Adaptation (UDA) has emerged as a solution. The goal of UDA is to align features (reduce domain gap) across the source (labelled) and target (unlabelled) domains, so that the adapted model performs well on both domains.

Another key issue at the production end is the ***memory constraint*** of the system. Most of the best-performing models are very deep and have high memory requirements. However, practical applications require the model to fit and evaluate on large images (as captured from a car) with limited memory, as in mobile devices. Many recent approaches have been inviting attention towards using compact light weight models [24],[45],[29]. However, these models result in lower performance efficiency in comparison with the state-of-the-art (SOTA) architectures that use deep backbones. With respect to this, model compression techniques like Knowledge Distillation [16],[32] have contributed significantly in fuelling the performance of a compact *Student* network by transferring knowledge from a cumbersome *Teacher* network.

In this paper, we propose a novel holistic end-to-end approach for distilling domain adaptive segmentation knowledge to tackle the challenges of domain gap and memory constraint, simultaneously. The goal is to distil domain-adaptive knowledge from a computationally expensive teacher network to a compact student network, to boost its performance in the target domain. To the best of our knowledge, this paper marks one of the first approaches to-

wards this multifaceted goal. To facilitate our proposal, we propose a *multi-level distillation* (MLD) strategy with novel loss functions to improve performance for both knowledge distillation (KD) and domain adaptation (DA) in a simultaneous manner.

Our proposed method distils knowledge at two levels - feature space and output space - using a combination of KL divergence and MSE loss terms. Furthermore, in order to optimize our model for both tasks simultaneously, we discern the need for an additional loss function that is tailored explicitly to serve the multitudinous goal. In this regard, we leverage the teacher network soft predictions to generate pseudo labels. These pseudo labels enhance knowledge transfer at the output space. These labels are particularly relevant in the target domain, as they serves as a proxy for ground truth, which is otherwise unavailable. Our newly presented cross entropy loss function with pseudo-labels serves as a multifarious objective function, by serving as an emissary for the unavailable ground truth in the target domain, while also amplifying the distillation from the teacher to the student.

We evaluate our proposed method on large-scale autonomous driving datasets. In addition to the benchmark synthetic-to-real adaptation case scenario (GTA5 to Cityscapes), we assess our pipeline on a real-to-real scenario too (BDD to Cityscapes). In summary, the key contributions of this paper are as follows:

(1) A novel framework that distils domain adaptive knowledge to address two core practical problems (i.e. low memory networks & domain gap) concurrently.; (2) A *multi-level distillation* strategy carried out at the feature and output spaces.; (3) Proposal of a tailor-made loss function: *cross entropy with pseudo teacher labels* for this multifaceted problem.; (4) Extensive experimental analysis and ablation studies on both *synthetic-to-real*, and on *real-to-real* cases.

2. Related work

2.1. Domain adaptation

Domain adaptation is a subfield of machine learning that deals with employing a model trained on a source distribution in the context of a different (but related) target distribution. The goal is to reduce the domain gap and to find a common feature space that works well across domains. This problem has seen immense progress in the past. Domain gaps can be reduced in an adversarial way [37, 36], by reconstruction [15, 26] or by minimizing distance between distributions [23, 33].

Domain-adaptive semantic segmentation is of keen interest to researchers as semantic labelling is an incredibly labor-intensive task. One of the first methods was reported in [17]. The method uses fully convolutional networks for semantic segmentation and adversarial training for domain alignment. Kalluri et al. [18] performs semi supervised domain adaptation by using an entropy module that minimizes a similarity metric computing by projecting the representation at each pixel to an embedding space of different directions. Murez et al. [26] uses cycleGAN to transform images from one domain to another, and then uses both the original and the transformed images to find a model agnostic feature space. Tsai et al. [35] uses multi level adversarial learning to align features at different levels of the network. Vu et al. [38] uses losses based on the entropy of pixel wise predictions. Chen et al. [7] uses an image-to-image translation network, domain-specific task networks and cross domain consistency to find a common feature space.

2.2. Model compression via knowledge distillation:

Recently, there has been a surge in papers addressing model compression techniques in various vision problems. A thorough review of various compression techniques is reported in the survey paper [8]. One of the pioneer works in the field by Hinton et al. [16] showed that the extra information a.k.a. "dark knowledge" encapsulated in the soft targets (class probabilities produced by teacher) could be used for training the compact student model. Motivated by this observation, several other works were followed leveraging more teacher information. For instance, Romero et al. [31] presented an idea for distilling from the intermediate feature representations, whereas Yim et al. [42] proposed to distil the knowledge in terms of flow between layers, which is calculated by computing the inner product between features from two layers. In addition to various KD models in image classification [6],[9], it has been successfully applied in object detection [4], video classification [3] and speech recognition [40] as well. The idea has been explored for semantic segmentation tasks as well. Ros et al. [32] investigated the problem of road scene semantic segmentation leveraging different knowledge transfer frameworks based on the output probability of the teacher network. Further work includes distilling zero-order and first-order knowledge from teacher models [41] and distilling structured knowledge via pairwise and holistic distillation [21]. The paper on feature level distillation [12] explores the topic in the setting of action recognition.

2.3. Differences from past methods

While we acknowledge that there has been good headway on the ideas of domain adaptation and knowledge distillation as two separate problems, it is notable that in this paper, we conceptualize the best of both in a single holistic model, in a novel way. To the best of our knowledge, this is for the first time that the usage of domain adapta-
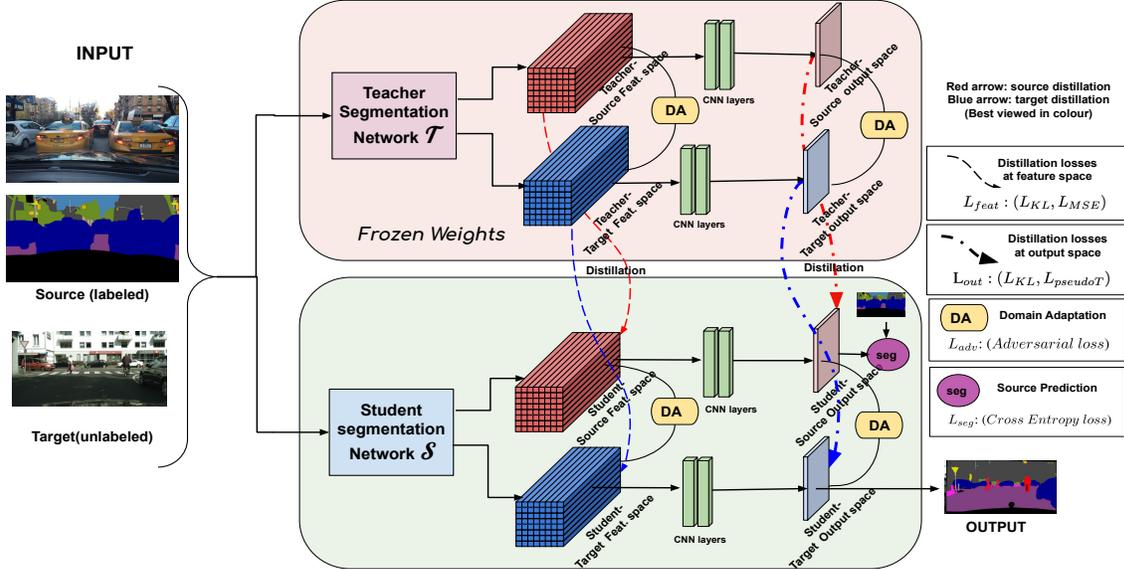
Figure 1: Our proposed framework for unsupervised domain adaptive knowledge distillation for driving scene semantic segmentation.

tion along with knowledge distillation has been explored in semantic segmentation. Albeit some early works addressed both of these concepts, those were on other problem settings such as Semi-supervised domain adaptation (SDA) [1],[28] or upon clinical/acoustic platforms [28],[2]. Further, they used knowledge distillation as a means for domain adaptation. On the contrary, we address a much more difficult problem setting - Unsupervised Domain Adaptation (UDA), on autonomous driving scene understanding, an area of stupendous practical implication. Unlike leveraging distillation as a means for domain adaptation [13],[27],[47], the design of our architecture is very well fine-tuned to serve both the goals of compression & domain adaptation in a specialized manner. Pseudo labels have been used for self-training in certain domain adaptation architectures [48, 10, 14, 25]. Additionally, noisy labels have also been used for partial knowledge transfer [20, 39]. In contrast, our paper proposes a method that leverages pseudo labels and improves both domain adaptation performance and knowledge distillation performance in a complementary manner.

3. Distillation model for UDA semantic segmentation

We propose a novel framework for distilling domain adaptive knowledge from a cumbersome teacher network to a lightweight student network, in the context of unsupervised semantic segmentation. In this section, the overall network architecture, distillation strategies and the associated objective functions are explained.

3.1. Overview of Network Architecture

The proposed network architecture is shown in Fig. 1. It consists of two neural networks *viz.* a cumbersome teacher segmentation network $\mathcal{T}$ and a compact student segmentation network $\mathcal{S}$. RGB images from both source and target domains and the segmentation maps for the corresponding source domain images are the inputs to the networks. The problem is completely unsupervised in the target domain *i.e.,* no labels are available for the target domain. First, the teacher $\mathcal{T}$ and the student $\mathcal{S}$ networks are initialized with parameters trained for the task of domain adaptation (DA) on a source-target dataset pair. The baseline DA model used in both student and teacher networks is detailed in later Section 3.2. The teacher network $\mathcal{T}$ (weights frozen after training) is used to distil knowledge to the compact student network $\mathcal{S}$ (distillations are shown in Fig. 1 via arrows). The task of the student network is to learn a domain-adaptive set of parameters, while taking advantage of the mastery of the teacher network. The transfer of knowledge can help the much smaller student network perform better than its innate capabilities.

In order to facilitate efficient knowledge distillation, we propose a multi-level distillation scheme, wherein knowledge transfer is carried out at multiple levels of the network: feature space and output space, as explained in Section 3.3. Further, two modes of distillations are designed at each level *i.e.* source distillation (red arrows) and target distillation (blue arrows). In addition, at the output space, we introduce a loss function that leverages pseudo labels from the teacher network. These pseudo labels serve two purposes: knowledge distillation from the teacher to the student & proxy for

the target domain ground truth, which is otherwise unavailable. A detailed explanation on the pseudo teacher label based distillation is provided in Section 3.3.3.

3.2. Baseline domain adaptation (DA) model

It is important to use a good domain adaptation strategy for each of the teacher and student networks. To this end, we use the popular multi-level domain adaptation strategy [35] to pretrain both the networks. To train the generator, segmentation loss L_{seg} (cross entropy loss) is applied on images from the source domain, and adversarial loss L_{adv} is applied for images from the target domain. The discriminators are optimised with a binary cross entropy loss applied on images from both the domains. The mathematical formulations [35] are as follows:

$$L_{seg} = -\sum_{h,w}\sum_{c\in C} Y_s^{(h,w,c)}\log(P_s^{(h,w,c)}) \qquad (1)$$

$$L_{adv} = -\sum_{h,w} log(D(P_t)^{h,w,1}) \qquad (2)$$

where Y_s represents ground truth, P_s is the probability map predictions for source images and P_t is the probability map predictions of target images. C represents the number of classes. The net equation to optimise the generator network is a sum of the segmentation and adversarial loss terms at feature and output level. This is optimized using the min-max criterion.

3.3. Multi-level distillation

We present a multi-level distillation method to transfer knowledge at both the feature and the output space to bestow the student network with improved performance capabilities while retaining its compactness. In this regard, we present the following loss functions:

3.3.1 KL divergence loss L_{KL}

The output of the networks are probability distributions. We use the KL divergence loss [19] to motivate the student to achieve distributions close to the teacher at the output level. This encourages the output of the student network to emulate the output of the teacher network.

$$L_{KL} = \lambda_{KL} * \sum_{i} KL(q_i^s||q_i^t) \qquad (3)$$

Here, λ_{KL} represents the weight hyperparameter for KL divergence distillation, q_i^s denotes student domain probability maps and q_i^t denotes the corresponding counterpart for teacher domain. To facilitate enhanced distillation, we enforce another KL divergence loss term at the high dimensional feature space (where feature-space adaptation

is done), nevertheless with a lower weight hyperparameter. Note that, in consistency with multi-level DA, Adapt-SegNet [35], we project the feature maps to segmentation probability maps. This enables us to apply KL divergence distillation at the feature space (Mathematically, KL divergence can be used to only constrain probability distributions). Thus, this loss is implemented at two stages viz., output and feature levels and we denote them as $L_{KL,out}$ and $L_{KL,feat}$ respectively. It is mathematically formulated as:

$$L_{KL} = L_{KL,out} + L_{KL,feat} \qquad (4)$$

3.3.2 MSE loss L_{MSE}

While the output space captures the crux of knowledge (scene layout, context, etc.) needed for distillation, the high dimensional feature space encodes complex representations and prove useful too. This leads us to impose a mean square (MSE) loss L_{MSE} between the feature maps of the two networks, which pushes them to resemble each other. KL divergence at the feature space aims to align probability distributions (soft alignment), while MSE pushes the feature spaces to exactly resemble each other (hard alignment). While MSE can promote the final layers of the 2 networks to exactly resemble each other, probability distributions can be adversely affected. KL divergence (log based loss) acts as a regularizer, while simultaneously aiding distillation. L_{MSE} is formally represented as:

$$L_{MSE} = \lambda_{MSE} * \sum_{i} ||(q_i^s - q_i^t)||^2 \qquad (5)$$

These feature maps correspond to the level at which feature-level domain adaptation is done. If the feature size between the student and teacher model is different, bilinear upsampling (proven to be effective in generating segmentation maps from feature maps) can be used to make the feature sizes same, followed by the application of the loss functions.

3.3.3 Domain-adaptive distillation with pseudo teacher labels $L_{pseudoT}$

This distillation happens at the output space, in addition to the distillation using KL-divergence loss. Based on the soft labels (probability maps) provided by the teacher, we determine the class that each pixel belongs to generate pseudo labels. Although these pseudo labels generated by the teacher are not totally accurate, the teacher network nevertheless performs better than the undistilled student network. These pseudo labels serve two purposes:

(i) Knowledge distillation for images in the source domain - The information from source domain pseudo-labels
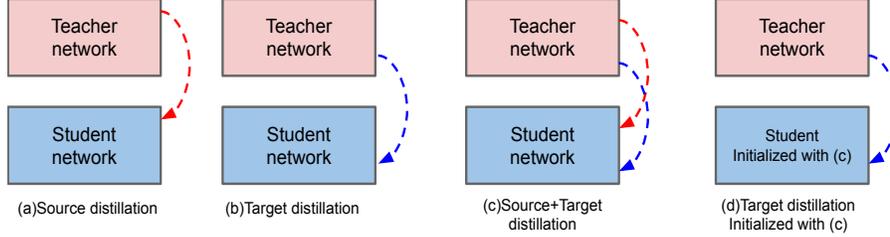
Figure 2: Various distillation paradigms. Source distillation and target distillation are shown in red and blue arrows respectively. (Best viewed in colour)

is distilled to the counterpart student network via multi-class cross entropy loss term. This is in addition to the segmentation multi-class cross entropy loss L_{seg} that is applied using the ground truth for images from the source domain. **(ii) Proxy ground truth for images in the target domain**, where the problem is completely unsupervised. The enchantment of these pseudo-labels is mainly exhibited in the case of the student target domain, wherein teacher domain pseudo-labels serve as a proxy for the ground truth. On this notion, we apply cross entropy loss between the student network target domain outputs and the corresponding pseudo-labels from teacher, that we term as $L_{pseudoT}$. While one may argue that for probability distributions, KL divergence and cross entropy play very similar roles and are thus equivalent to each other, we wish to emphasize that KL divergence is applied between the teacher and student probability distributions wherein the teacher distribution carries information about all classes. This can thus capture knowledge about semantically similar classes. However, for the pseudo label cross entropy loss term, we deterministically predict the label for each pixel from the teacher network probability distribution. This carries no information about the labels for other classes, and can also introduce false positives, which is countered with the KL divergence loss term. We observe that these pseudo-labels significantly improve the results. Mathematically,

$$L_{pseudoT} = \lambda_{pseudoT} * \sum_{h,w} \sum_{c} Y_{pseudoT} \log(P_{student})$$

(6)

$\lambda_{pseudoT}$ represents the hyperparameter for distillation with pseudo-labels, $Y_{pseudoT}$ denotes pseudo-labels from the teacher network, and $P_{student}$ denotes predictions (probability map) of the student network. The equation takes the same form for both source and target domain images.
Overall objective function:

The student network is trained with both domain adaptation losses (section 3.2), and distillation losses. The overall objective function for distillation is the weighted sum of the individual distillation loss functions. Taking advantage of the similitude between the source (s) and target (t) domains, we propose to introduce a scaling factor λ_{target}. For target domain distillation, we use parameters that are λ_{target}

times the corresponding parameters used for source domain distillation.

$$L_{distill} = [L_{s-KL} + L_{s-MSE} + L_{s-pseudoT}]+ \\ \lambda_{target} * [L_{t-KL} + L_{t-MSE} + L_{t-pseudoT}].$$

(7)

3.4. Distillation paradigms

We institute four distillation paradigms (Fig. 2). Fig. 2(a) represents the case where only images from the source domain are distilled. Similarly, Fig. 2(b) represents the case where only images from the target domain are distilled. Fig. 2(c) represents source + target distillation, the complete model as proposed in Fig. 1. Fig. 2(d) represents target distillation, where the student network is initialised with weights trained as in case 2(c). Each of these distillation paradigms contributes to improving performance of the overall architecture in different aspects, and we discuss these effects in detail in section 5.1.

4. Experimental Setup

4.1. Datasets & Evaluation Protocols

We evaluate the proposed algorithm on three popular large-scale autonomous driving datasets. The **Cityscapes (CS)** [11] dataset and **Berkeley Deep Drive (BDD)** [44] dataset are real scenes captured in Europe and the USA respectively. In particulat, BDD captures varying illumination conditions, seasonal changes. **GTA5** [30] is a popular synthetic driving dataset, and majorly emerged out of computer games. For Cityscapes, we use only the finely annotated images. The class definitions in all these datasets are compatible with each other.

We evaluate our models on the validation images of the target domain dataset, using the standard segmentation evaluation metrics: Intersection over Union (**IoU**) and pixel wise accuracy. The IoU score is calculated as the ratio of intersection and union of the ground truth mask and the predicted mask for each class. We report classwise IoUs for and also compute the **mIoU** (mean IoU) as the mean of all classwise IoU scores. Additionally, we report the overall **pixel accuracy** as the ratio of the pixels classified correctly to the number of overall pixels.

Experiment	mIoU	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	MBike	Bike	mAcc
							(I). Results on the real-to-real scenario: BDD to Cityscapes														
$\mathcal{T}$: DRN-D-38	**42.33**	91.53	52.72	80.65	19.21	23.4	26.31	23.75	35.79	82.48	33.85	75.36	47.87	17.65	83.15	33.77	36.74	0.08	11.02	29.01	**88.03**
$\mathcal{S}$: DRN-D-22	**38.33**	89.25	49.03	78.2	16.22	20.27	21.17	17.28	33.51	80.8	28.89	71.97	45.44	12.7	76.1	21.35	32.18	0.91	6.35	26.66	**86.12**
Distillation (a)	**42.33**	92.41	53.91	80.57	19.3	22.89	26.88	23.03	36.05	82.59	34.67	77.42	48.39	15.39	82.8	27.57	40.04	0.03	9.22	31.12	**88.42**
Distillation (b)	**43.73**	92.65	56.1	81.55	17.04	26.65	26.1	25.66	39.36	83.61	37.4	76.29	50.11	14.41	83.6	33.25	41.03	0.34	10.64	35.07	**88.89**
Distillation (c)	**43.97**	92.73	56.51	81.57	16.34	26.21	27.18	25.28	38.47	83.96	37.74	79.17	50.78	15.82	83.67	33.59	40.95	0.08	10.03	35.36	**89.09**
Distillation (d)	**44.15**	92.64	55.91	81.74	16.3	26.53	26.51	25.88	38.72	83.97	37.27	78.61	50.77	16.13	83.94	36.14	39.87	0.28	11.35	36.28	**89.08**
							(II). Results on the synthetic-to-real scenario: GTA to Cityscapes														
$\mathcal{T}$: DRN-D-38	**32.49**	87.02	40.43	73.12	17.29	10.85	21.25	19.32	10.69	71.88	26.1	58.64	41.34	9.26	75.84	15.87	21.26	10.12	5.86	1.13	**82.23**
$\mathcal{S}$: DRN-D-22	**30.39**	86.88	38.49	69.92	13.98	12	19.73	18.05	10.09	69.09	25.39	51.06	39.64	4.93	73.5	12.71	17.36	5.6	6.59	2.44	**80.83**
Distillation (a)	**31.91**	87.59	39.42	72.19	15.99	11.16	20.26	20.46	11.89	71.76	26.93	60.5	41.49	4.33	76.21	14.19	18.89	5.48	6.3	1.68	**82.63**
Distillation (b)	**32.98**	88.76	43.5	73.34	18.56	11.32	20.39	18.2	11.5	72.67	29.45	59.47	43.35	6.65	78.53	17.24	23.71	3.48	4.97	1.5	**83.51**
Distillation (c)	**33.36**	88.88	43.99	73.69	19.88	11.37	20.91	19.18	10.94	73.16	29.57	59.37	43.43	7.5	78.83	17.9	23.07	4.94	5.83	1.39	**83.72**
Distillation (d)	**33.81**	89.07	44.84	74.15	20.03	11.41	22.06	19.19	11.83	73.58	29.58	59.9	43.59	7.64	79.12	18.13	22.09	8.18	6.47	1.45	**83.9**

Table 1: Results of our proposed Domain Adaptive Knowledge Distillation method on synthetic-to-real and real-to-real cases scenarios

4.2. Implementation details

We follow the standard procedure for training a GAN and alternate between two steps: i) Fix the discriminator and train the segmentation network with segmentation loss (source domain), domain discrimination loss (target domain) and distillation losses (source and target domains) ii) Fix the generator network and train the discriminator. Details regarding the learning rate, optimizers and GPU configurations are provided in the supplementary material.

5. Experimental results

The performance of our model is evaluated in various distillation paradigms on real-to-real and synthetic-to-real scenarios. Extensive ablation studies are also carried out to determine the performance of various components of our model under various settings.

5.1. Real to real adaptation: Berkeley Deep Drive to Cityscapes

The results of domain-adaptive distillation from a real environment to another real environment viz., BDD to CS are shown in Table 1(I) (Visual results in Fig. 3). The first two rows represent the baseline teacher(DRN-D-38, 26.5 M parameters) and student(DRN-D-22, 16.4 M parameters) networks respectively, that are trained using multi-level DA [35]. The results in the subsequent four rows correspond to distillation via the four distillation paradigms (section 3.4). The weights for knowledge distillation (in both source and target domains) are set at 0.1, 0.01 and 1.0 for λ_{KL} (output level), λ_{MSE} (feature level) and $\lambda_{pseudoT}$ (output level) respectively. For KL divergence distillation at the feature level, we set the hyperparameter as one-tenth of its counterpart at the output level.

Analysis of the effects of distilling via the four distillation paradigms (as per Fig 2): As we can observe in Table 1, source distillation (case (a)) and target distillation (case (b)) outperform the corresponding undistilled student network by 10.43 % and 14.08 % (4 and 5.4 mIoU points

in terms of absolute numbers) respectively. Target distillation performs better than source distillation. This can be ascribable to the fact that distilling only in the target domain reduces the bias of the model towards the source domain, and improves performance in the target domain. In addition, our network in case (b) (target distillation) outperforms even the teacher network by 3.30 % (1.4 mIoU points in terms of absolute numbers). Further, in concurrence with our intuition, case (c) (where both source and target distillation are done simultaneously) performs better than both case (a) and case (b), achieving an mIoU of 43.97 (14.71 % and 3.87 % relative improvement over the student and teacher networks respectively). Among all the distillation case studies, we observe that case (d) performs the best with a improvement of 15.18 % and 4.29 % (5.82 and 1.82 mIoU points respectively) over the student and teacher networks respectively. This is because the model in this case has the best of both worlds - it is initialised with the weights of case (c) thus providing it with the pre-requisite information necessary for source distillation. Finetuning it with target distillation brings target distillation at the focus of the model thus improves the results in the target domain via generalization. We also accredit the improved performance gain of the student network via knowledge distillation to the customized objective functions that we employed in this task, such as cross-entropy loss with pseudo teacher labels $L_{pseudoT}$. We conjecture that the dark knowledge [16] encapsulated within these pseudo-labels in both source and the target domains substantially augment the individual performance of the compact student network. The fact that our network surpasses even the teacher network by 4.29 % (1.82 mIoU points in terms of absolute numbers), which is an appreciable number, is commendable given that the teacher network has almost 10 M parameters more than the student network. We notice similar trends in the mAcc scores too.

5.2. Synthetic to real adaptation: GTA5 to Cityscapes

Experiments on the synthetic to real dataset pair: GTA5 to Cityscapes with DRN-D-38 teacher (26.5 M parameters) and DRN-D-22 student (16.4 M parameters) are presented in Table 1(II) (Visual results in Fig. 3). The trends are analogous to the real to real adaptation case. Source distillation (case (a)) shows an improvement of 5 % over the undistilled student network (1.52 mIoU points). The other distillation paradigms (case b,c,d) not just outperform the undistilled student network by 8.52 % - 11.25 % (1.52 - 3.42 mIoU points in terms of absolute numbers), but also the teacher model by 1.50 % - 4.06 % (0.49 - 1.32 mIoU points in terms of absolute numbers). This improvement can be explained on the basis of the knowledge transfer from the teacher to student.

Effect of change of teacher and student networks: It is important to study the effects of change of teacher and student networks in knowledge distillation algorithms (Table 2(a)). In addition to the baseline DRN-D-38 teacher & DRN-D-22 student networks, we use DeepLab (ResNet-101, 44.5 M parameters) as the teacher network and experiment with DRN-D-38 (26.5 M parameters) and DRN-D-22 (16.4 M parameters) as the student networks, with [35] baseline architectures. Our distilled student networks demonstrate an improvement of 13.14 % - 19.11 % (4.27 - 5.81 mIoU points in terms of absolute numbers) over the corresponding undistilled student networks, thus showing the effect of knowledge transfer from a powerful and heavy teacher network to different lightweight student networks.

Effect of change of Domain Adaptation backbone: Since our pipeline is immune to the internal domain adaptation strategy used in the teacher and student networks, it can be treated as a blackbox and can be replaced by any other powerful DA approach. Our domain adaptive distillation pipeline will help the student network to benefit accordingly. Table 2(b) shows the effect of using the domain adaptation method, ADVENT [38] with DeepLab (ResNet-101, 44.5 M parameters) backbone, as the teacher. We demonstrate an improvement of 20.63 % and 14.9 % (6.27 and 4.86 mIoU points respectively) over the corresponding undistilled DRN-D-22 and DRN-D-38 student networks respectively.

SOTA comparisons: Since this paper marks the first approach towards this goal of tackling both domain adaptation and model compression issues in a simultaneous manner, the results shown in this paper set the benchmark for the problem. In addition, in the interest of thorough analysis, we compare the performance of our student model on the synthetic-to-real case against the performance of the current SOTA UDA methods on models with similar sizes, for fair comparison. The results are shown in Table 3. With the same backbone architecture (DRN-D-22, first row)

and baseline DA strategy, our best model outperforms the popular multi-source DA method AdaptSegNet [35] by 4-6 mIoU points (14%-20%) with a model size that is $0.6x$ the model size of its counterpart. By extension, our model scales appropriately when evaluated against all other methods that have similar model sizes and are trained using similar DA strategies. This shows that our method enables our relatively lightweight model achieve SOTA results against comparatively heavier models, which is a testimony to the success of our proposed method.

Network	mIoU	Params(M)
$\mathcal{T}$: DRN-D-38	32.49	26.5
$\mathcal{S}$: DRN-D-22(No dist.)	30.39	16.4
$\mathcal{S}$: DRN-D-22(DA dist. (Ours))	**33.36**	16.4
$\mathcal{T}$: DeepLab	42.5	44.5
$\mathcal{S}$: DRN-D-38 (No dist.)	32.49	26.5
$\mathcal{S}$: DRN-D-38 (Ours)	**36.76**	26.5
$\mathcal{T}$: DeepLab	42.5	44.5
$\mathcal{S}$: DRN-D-22 (No dist.)	30.39	16.4
$\mathcal{S}$: DRN-D-22 (Ours)	**36.2**	16.4

(a) Varying teacher($\mathcal{T}$) and Student($\mathcal{S}$) networks

Network	mIoU	Params(M)
$\mathcal{T}$ and $\mathcal{S}$ DA:AdaptSegnet [35]		
$\mathcal{T}$: DRN-D-38	32.49	26.5
$\mathcal{S}$: DRN-D-22 (Ours)	**33.36**	16.4
$\mathcal{T}$ DA:ADVENT [38]; $\mathcal{S}$ DA: AdaptSegNet [35]		
$\mathcal{T}$: DeepLab	42.5	44.5
$\mathcal{S}$: DRN-D-22 (No dist.)	30.39	16.4
$\mathcal{S}$: DRN-D-22 (Ours)	**36.66**	16.4
$\mathcal{T}$ DA:ADVENT [38]; $\mathcal{S}$ DA: AdaptSegNet [35]		
$\mathcal{T}$: DeepLab	42.5	44.5
$\mathcal{S}$: DRN-D-38 (No dist.)	32.49	26.5
$\mathcal{S}$: DRN-D-38 (Ours)	**37.35**	26.5

(b) Varying DA backbone

Table 2: Experiments by varying backbone networks

Method	Model	Model Size	mIoU
AdaptSegNet [35]	DRN-D-22	16.4	30.39
AdaptSegNet [35]	DRN-D-38	26.5	32.49
Ours ([35] backbone)	DRN-D-22	16.4	36.66
Ours ([35] backbone)	DRN-D-38	26.5	37.35

Table 3: State-Of-The-Art (SOTA) comparisons

5.3. Ablation Studies

To deduce the performance of the individual distillation loss functions, we conduct extensive ablation studies on the real-to-real adaptation case. We conduct these studies on distillation paradigm case (a) of Fig. 2. The rationale is that these results should scale for the other three cases as well.

5.3.1 Impact of various loss functions:

The impact of various distillation loss functions (which also correspond to studies on feature level distillation vs output level distillation) are shown in Table 4. For each of the
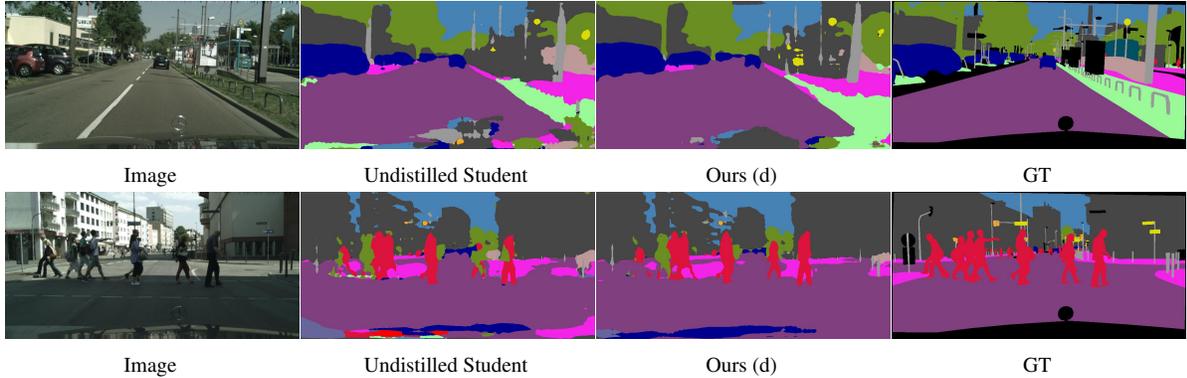
Figure 3: Visual results - First row: BDD to Cityscapes and Second Row: GTA5 to Cityscapes (More results can be found in the supplementary material). Ours (d) corresponds to distillation paradigm (d) as per Fig. 2

distillation loss terms i.e., L_{KL}, L_{MSE} and $L_{pseudoT}$, we conduct ablation studies by varying the corresponding hyper parameters (see Table 4(i),(ii) and (iii)). The ablation studies on loss term L_{MSE} and $L_{CE-pseudo}$ are correspondent with feature level and output level distillation respectively. L_{KL} is employed according to the MLD strategy. We observe that the loss function $L_{pseudoT}$ performs better than both L_{KL} and L_{MSE}, by achieving the best mIoU values of 40.27, 39.86 and 42.18 respectively. We observe that L_{KL} performs its best at a weight of 0.4 at the output level(feature level:0.04). Similarly, L_{MSE} performs its best at 0.01 and $L_{pseudoT}$ performs its best when its weight is 1. For pixels where the label (produced by the teacher network) is faithful, this loss function should mimic the ideal case of classical cross entropy loss function. A high hyperparameter (1.0) pushes the objective function towards this goal.

We also conduct experiments on various permutations of the three distillation loss functions, as per MLD scheme (see Table 4(iv)-(vii)). We observe that with careful tuning, each of the combinations can achieve better results than the individual loss terms involved. Our best results are obtained (mIoU 42.33) when we use all the three.

5.3.2 Impact of λ_{target}

(Table 4(viii)), ablation studies case (case Fig. 2(c))): The source distillation parameters are set at 0.1, 0.01 and 1.0 for the three loss terms respectively. For target distillation, we use the same parameters scaled by a factor λ_{target} (see equation 7). The best results are obtained at a value of 1.0 (shown in bold). This can attributed to the same reason of ideal scenario where 1.0 works best for $\lambda_{pseudoT}$. These teacher labels basically serve as a representative for target domain GT, where our problem is completely unsupervised.

6. Conclusion and Future work

We proposed a novel method for distilling domain adaptive segmentation knowledge in models with limited memory, which has important implications in autonomous driving scenarios. To the best of our knowledge, this paper

Parameter(λ)	mIoU	mAcc
(i) L_{KL}, Distillation paradigm (a)		
0.1	39.54	86.65
0.4	**40.27**	**87.15**
0.7	40.0	86.81
1.0	38.94	86.18
(ii) L_{MSE}, Distillation paradigm (a)		
0.005	38.91	86.25
0.05	38.71	86.44
0.01	**39.86**	**87.17**
(iii) $L_{pseudoT}$, Distillation paradigm (a)		
0.001	40.15	87.05
0.01	40.6	87.22
0.05	40.72	87.4
0.1	41.01	87.59
0.5	41.45	87.98
1.0	**42.18**	**88.28**
(iv) $L_{KL}+L_{MSE}$, Distillation paradigm (a)		
0.4, 0.01	39.89	87.15
0.1, 0.01	**40.32**	**87.37**
0.7, 0.05	39.7	87.17
(v)$L_{KL} + L_{pseudoT}$, Distillation paradigm (a)		
0.1, 0.1	41.18	87.98
0.1, 1	**41.92**	**88.3**
(vi) $L_{MSE} + L_{pseudoT}$, Distillation paradigm (a)		
0.01, 1	**42.2**	**88.54**
(vii)$L_{KL} + L_{MSE} + L_{pseudoT}$, Distillation paradigm (a)		
Case (a): **0.1, 0.01, 1**	**42.33**	**88.42**
(viii) λ_{target}, Distillation paradigm (c)		
0.05	43.22	88.76
0.1	43.05	88.67
0.5	43.64	88.9
1.0	**43.97**	**89.09**

Table 4: Ablation studies: Impact of various distillation losses on the real-to-real case, table 1I; Class-wise performance statistics can be found in the supplementary

marks the first approach towards this goal. In this regard, we presented a multi-level distillation strategy and proposed a tailor-made multifarious loss function to simultaneously handle two key issues in segmentation models, domain gap and memory constraints. We instituted various distillation paradigms. Our experiments and extensive analysis show the success of our proposed method.

References

[1] Shuang Ao, Xiang Li, and Charles X Ling. Fast generalized distillation for semi-supervised domain adaptation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 3

[2] Taichi Asami, Ryo Masumura, Yoshikazu Yamaguchi, Hirokazu Masataki, and Yushi Aono. Domain adaptation of dnn acoustic models using knowledge distillation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5185–5189. IEEE, 2017. 3

[3] Shweta Bhardwaj, Mukundhan Srinivasan, and Mitesh M Khapra. Efficient video classification using fewer frames. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 354–363, 2019. 2

[4] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems*, pages 742–751, 2017. 2

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[6] Wei-Chun Chen, Chia-Che Chang, and Che-Rung Lee. Knowledge distillation with feature maps for image classification. In *Asian Conference on Computer Vision*, pages 200–215. Springer, 2018. 2

[7] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1791–1800, 2019. 2

[8] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282*, 2017. 2

[9] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4794–4802, 2019. 2

[10] Jaehoon Choi, Minki Jeong, Taekyung Kim, and Changick Kim. Pseudo-labeling curriculum for unsupervised domain adaptation. *arXiv preprint arXiv:1908.00262*, 2019. 3

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5

[12] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2019. 2

[13] Zhijie Deng, Yucen Luo, and Jun Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9944–9953, 2019. 3

[14] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020. 3

[15] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016. 2

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 6

[17] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 2

[18] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and CV Jawahar. Universal semi-supervised semantic segmentation. *arXiv preprint arXiv:1811.10323*, 2018. 2

[19] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 4

[20] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1910–1918, 2017. 3

[21] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019. 2

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[23] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015. 2

[24] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–568, 2018. 1

[25] Pietro Morerio, Riccardo Volpi, Ruggero Ragonesi, and Vittorio Murino. Generative pseudo-label refinement for unsupervised domain adaptation. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 3130–3139, 2020. 3

[26] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018. 2

[27] Jogendra Nath Kundu, Nishank Lakkakula, and R Venkatesh Babu. Um-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation. *arXiv preprint arXiv:1908.03884*, 2019. 3

[28] Mauricio Orbes-Arteainst, Jorge Cardoso, Lauge Sørensen, Christian Igel, Sebastien Ourselin, Marc Modat, Mads

Nielsen, and Akshay Pai. Knowledge distillation for semi-supervised domain adaptation. In *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging*, pages 68–76. Springer, 2019. 3

[29] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016. 1

[30] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 5

[31] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2

[32] German Ros, Simon Stent, Pablo F Alcantarilla, and Tomoki Watanabe. Training constrained deconvolutional networks for road scene semantic segmentation. *arXiv preprint arXiv:1604.01545*, 2016. 1, 2

[33] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016. 2

[34] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5229–5238, 2019. 1

[35] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 2, 4, 6, 7

[36] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017. 2

[37] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 2

[38] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 2, 7

[39] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Kdgan: Knowledge distillation with generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 775–786, 2018. 3

[40] Jeremy HM Wong and Mark Gales. Sequence student-teacher training of deep neural networks. 2016. 2

[41] Jiafeng Xie, Bing Shuai, Jian-Fang Hu, Jingyang Lin, and Wei-Shi Zheng. Improving fast segmentation with teacher-student learning. *arXiv preprint arXiv:1810.08476*, 2018. 2

[42] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network

minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017. 2

[43] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. 1

[44] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018. 5

[45] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–420, 2018. 1

[46] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1

[47] Sicheng Zhao, Guangzhi Wang, Shanghang Zhang, Yang Gu, Yaxian Li, Zhichao Song, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source distilling domain adaptation. *arXiv preprint arXiv:1911.11554*, 2019. 3

[48] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 3